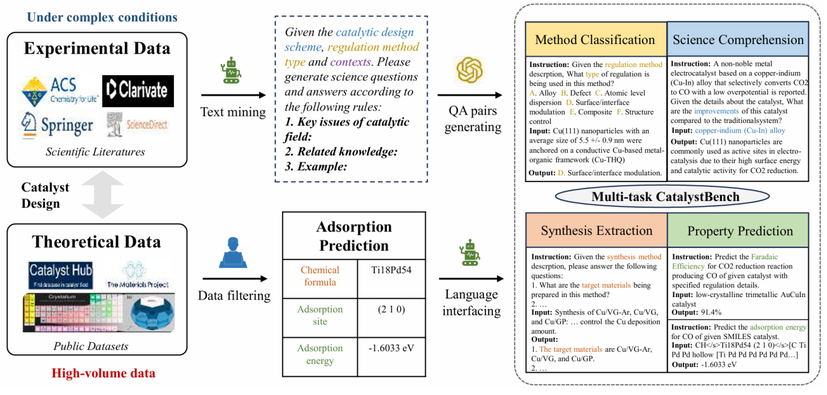
CatalystBench: A Comprehensive Multi-task Benchmark for Advancing Language Models in Catalysis Science

Xueqing Chen, Jian Xu, Ludi Wang, Yang Gao, Huihan Zhu, Yuanchun Zhou, Yi Du*, and Cheng-Lin Liu*

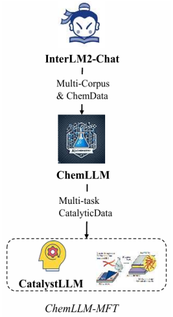
中国科学院 计算机网络信息中心 Computer Network Information Center, Chinese Academy of Sciences

国家纳米科学中心 National Center for Nanoscience and Technology

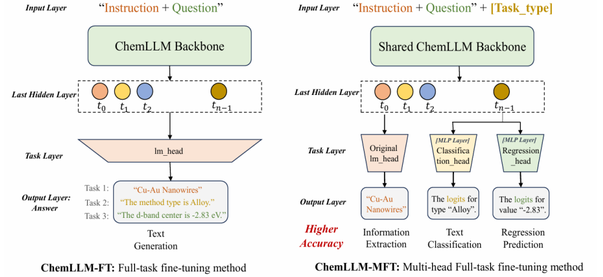
CatalystBench: The Benchmark



CatalystLLM: The Model



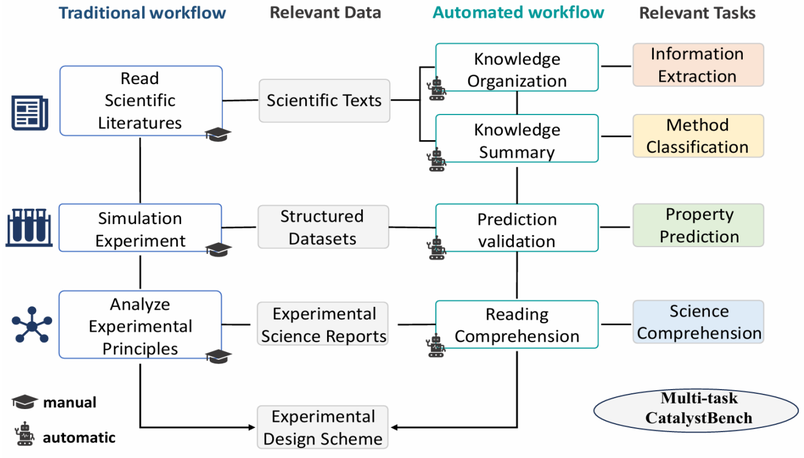
(a) Overview of CatalystLLM.



(b) Comparison of model architectures between ChemLLM-FT and ChemLLM-MFT method.

Motivation

*The benchmark should combine **theoretical simulations** with **scientific experimental data** in catalysis field.*



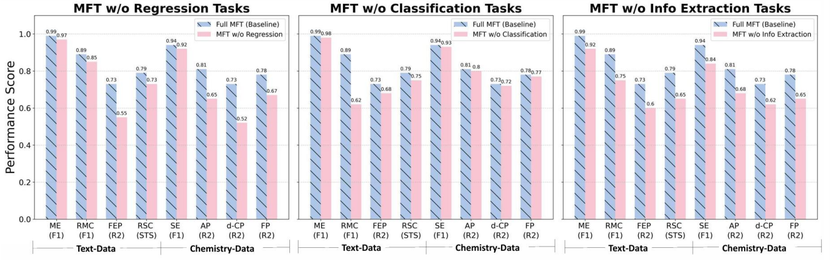
Main Contributions

- **CatalystBench**, a multi-task benchmark for catalysis, unifying theoretical data and experimental literature in one workflow-oriented framework.
- **MFT**, a multi-head full-task fine-tuning strategy that separates heterogeneous outputs and improves multi-task learning.
- **CatalystLLM** achieves sota performance and reveals both the strengths and limits of LLMs in catalysis.

Results

Model	ME		SE		RMC		FEP		AP		d-CP		FP	
	ACC	F1	ACC	F1	ACC	F1	R2	MAE	R2	MAE	R2	MAE	R2	MAE
Closed-source LLMs														
claude-3	0.93	0.95	0.84	0.90	0.58	0.72	0.30	6.73	0.37	3.58	0.33	4.21	0.39	5.24
Gemini-2.5	0.97	0.98	0.87	0.91	0.66	0.80	0.29	3.19	0.34	4.15	0.36	5.78	0.40	4.98
gpt-3.5	0.89	0.93	0.78	0.86	0.49	0.62	0.28	5.25	0.35	4.20	0.34	4.50	0.38	5.41
gpt-4o	0.96	0.98	0.88	0.92	0.69	0.81	0.36	3.10	0.43	3.05	0.41	3.35	0.46	4.02
gpt-4.1	**0.98**	**0.99**	**0.91**	**0.95**	0.75	0.85	0.56	2.51	0.61	2.40	0.59	2.15	0.65	2.68
Open-source LLMs														
deepseek-v3	0.92	0.95	0.83	0.89	0.61	0.74	0.33	4.20	0.40	3.55	0.38	3.85	0.43	4.45
LLaMA2-7B	0.83	0.86	0.73	0.79	0.39	0.50	0.25	4.47	0.24	4.63	0.34	3.79	0.35	4.12
Qwen3-8B	0.86	0.91	0.74	0.83	0.42	0.54	0.26	4.59	0.31	3.85	0.32	3.49	0.36	4.26
Mistral-7B	0.88	0.92	0.76	0.84	0.44	0.57	0.29	3.09	0.37	3.49	0.36	4.14	0.39	5.15
ChatGLM3-6B	0.84	0.90	0.74	0.83	0.45	0.57	0.33	4.87	0.39	3.07	0.37	3.48	0.41	4.38
ChemLLM	0.93	0.95	0.79	0.88	0.52	0.66	0.45	2.80	0.63	2.05	0.54	2.36	0.64	2.75
Darwin1.5	0.91	0.94	0.79	0.89	0.50	0.64	0.44	2.81	0.59	3.13	0.54	2.42	0.68	2.01
CatalystLLM	**0.98**	**0.99**	0.89	0.94	**0.81**	**0.89**	**0.73**	**1.72**	**0.81**	**1.24**	**0.73**	**1.49**	**0.80**	**1.34**

- CatalystLLM ranks first **overall** on factual-answer tasks.
- **Specialized models** outperform general LLMs on numerical regression tasks.
- **General LLMs** remain competitive on ME, SE and RMC.



- **Intra-group** synergy is strong: tasks with similar input types help each other a lot.
- There is also **inter-group** synergy between text tasks and chemical data tasks.